The use and usability of inferential techniques

Hoekstra, Rink

5. Interpreting Research Results on the Basis of *p*-value and Sample Size

Abstract

Researchers, especially in the behavioural sciences, usually report the *p*-values associated with their test statistics. However, what they would really like to know, it could be argued, are the probability that an effect exists in the population and the probability that an attempt to replicate will give a similar result. In our study, we asked 51 university students and lecturers to estimate these probabilities after reading a short summary of experiment outcomes, with only *p*-value and sample size given. Although large individual differences were found, we also found for a large proportion of the university students and lecturers that their estimated probabilities increased with an increase of sample size and fixed *p*-value. Numerical Bayesian assessments of these probabilities, assuming a uniform prior for the difference in population means, showed this direction of estimates to be incorrect for both probabilities.

5.1 Introduction

For most studies in the behavioural sciences it is impossible to study the complete population, and therefore studies are based on samples. Statistics is used to draw inferences from the sample data to the population one is interested in. The most frequently used statistical method for this purpose is *null hypothesis significance testing* (NHST), despite all the criticism it has received, as has extensively been described in Chapter 1. In the present Chapter, it is studied how the NHST outcomes are typically interpreted.

NHST is designed to assess the strength of the evidence against a null hypothesis (H_0). Typically, the null hypothesis is that there is no effect

in the population. The so-called *p*-value is used to assess the strength of evidence against H_0. It is the probability of finding a test statistic with a value as extreme as or more extreme than the observed value, assuming that the null hypothesis is true. The smaller this *p*-value, the stronger the evidence against H_0 (e.g., Moore & McCabe, 2003). Note, however, that this is not what has always been taught. In earlier introductory statistics books, inaccurate definitions could often be found (see, e.g., Gigerenzer, Krauss & Vitouch 2004, for a discussion). For example, *p*-values smaller than the adopted significance level have been defined as indicating an effect, and larger *p*-values were as indicating the absence of an effect. Although the *p*-value may be accompanied by other statistics, such as effect size, this is often not the case in practice (Finch, Cumming & Thomason, 2001). Furthermore, it is typically the *p*-value, or whether or not the *p*-value is significant, that is used to come to a conclusion about the obtained results (Finch et al.), thus stressing the central position of the *p*-value in inferential statistics.

Given this central position of the *p*-value in inferential statistics, one might expect that researchers agree on how to interpret the *p*-value. As is shown in Chapter 1, often errors are made in interpreting the outcomes of a significance test. The results suggest that researchers use NHST outcomes to answer questions that cannot be answered by those outcomes alone. *P*-values, in particular, are given incorrect interpretations.

There seem to be two important probabilities that researchers would like to be able to estimate. Because the researcher is primarily interested in the population and not in the sample itself, a first crucial probability is the probability that an effect in the same direction or of similar magnitude is present in the population. This generalizability of the results seems a central concept in inference, but there is another concept that is often regarded important as well: The replicability of the study, defined here as the

probability that a replication study will show a significant effect in the same direction. Nearly half a century ago the crucialness of these two probabilities was mentioned by Lubin (1957), who wrote that "assuredly all editors employ [replicability and generalizability] in judging the soundness of an article" (p.519).

We refer to the probability that an effect in the same direction or of similar magnitude is present in the population as the "certainty probability" (or, "certainty", for short), because it refers to the certainty of the existence of an effect in the population. When the difference between two means is of interest, we define certainty as the probability that the difference in means in the population is in the expected (as defined in advance in the alternative hypothesis) direction.

The second probability of interest, the probability that a replication study will show a significant effect in the same direction as the original study we call the "replicability probability" (or, "replicability", for short). Here, a "replication study" is operationalised as a replication of the initial study using the same variables with the same sample size, with the sample drawn in the same way from the same population (Posavac, 2002).

Critically, neither probability can, without further assumptions, be directly related to the *p*-value (see Appendix). However, it seems unlikely that significance testing could have achieved its central place in psychological research without an implicit connection between *p*-values and these probabilities. Indeed, *p*-values are often incorrectly interpreted as the complement for certainty, and as replicability (Kirk, 1996). Given the importance of both probabilities for researchers, and given the central role of *p*-values in inferential studies, researchers can reasonably be expected to have at least some intuitive estimate of both probabilities when interpreting research outcomes.

Apart from *p*-values, things that may influence the conclusions drawn from experimental results include sample sizes, confidence intervals, and effect sizes. In practice (e.g., Finch, 2001; Hoekstra et al., 2006), however, confidence intervals and standardised effect sizes are rarely reported, leaving only sample size (*n*) as a statistic that may influence the interpretation of *p*-values.

Even when *n* is known it is still not possible to compute the certainty and replicability probabilities, without further assumptions. However, under some simple assumptions about the distribution of the population effect size the nature of change of the probability estimates can be assessed, as described in the appendix. When *n* is fixed and the *p*-value decreases (and all else remains the same), the directions of change of both certainty and replicability are such that as *p* becomes smaller, both probabilities increase. Conversely, when the *p*-value is fixed and *n* increases, both probabilities are virtually independent of *n*.

In the present study, our primary interest was how researchers estimate certainty and replicability in practice. As *p*-values are, in practice, an important way to communicate inferential outcomes, one might expect intuitive estimates of certainty and replicability based on them to be more or less similar across researchers. Furthermore, we were interested in whether the conclusions researchers draw from the comparison of statements with the same *n*, but different *p*-values, and vice versa, are in line with the outcomes of our assessments of the relations of certainty and replicability to *n* and *p*-value.

Some previous evidence that people may incorrectly formulate certainty probabilities comes from a study by Rosenthal and Gaito (1963), in which they asked psychologists to “rate their degree of belief or confidence in a variety of *p* levels”, assuming different values for *n*, on a five-level scale. This degree of belief can be interpreted as the degree of belief of the

existence of an effect in the expected direction in the population, or, in other words, the certainty probability. Rosenthal and Gaito found that the psychologists indicated having more confidence when the sample size increased, even though the *p*-value stayed the same. This result suggests that intuitive ideas about the role of *n* may be incorrect.

With regard to the replicability probability, Tversky and Kahneman (1971) showed that people's intuitions about random sampling may affect their judgments. They asked people to estimate the probability that a study with a given sample size and given *z*-value or *t*-value will replicate when the replication study is carried out with a smaller *n*. They found that subjects overestimated the replicability probability, when compared to outcomes calculated by means of Bayesian reasoning using a uniform prior. According to Tversky and Kahneman, the subjects showed this behaviour because people tend to overestimate the representativeness of samples. Because of the limited number of *z*-and *t*-values that were presented in Tversky and Kahneman's study, it is hard to generalize their results to the way *n* and *p*-value influence replicability.

In the present study, we examine directly the certainty and replicability probability estimates people make when confronted with a range of sample sizes and *p*-values. These estimates are compared to mathematically derived relations and simulation study results to examine the correctness of the judgments. Furthermore, intersubject variability is examined in detail to determine the boundary condition for communicating certainty and replicability probabilities.

5.2 Experiment

We conducted an experiment in order to address whether researchers' estimates of certainty and replicability differ as a function of *n* and *p*-value and whether these estimates are in line with our numerical

assessments of this. Groups of undergraduate students, Ph.D. students and psychology lecturers were asked to read short scenarios in which only *n* and *p*-values were given and to make judgments about how certain they were that the effect was present in the population (the certainty probability) or how certain they were that the effect would be significant with a new sample of the same size (the replicability probability).

Method

Subjects. Fifty-one subjects, aged 19-60 years (mean = 31.0, SD = 11.5; 29 women) took part in the study. The group consisted of three subgroups, with varying statistical data-analysis expertise: undergraduate students, Ph.D. students, and lecturers, all from the psychology department at the University of Groningen. The undergraduate students had all attended at least three introductory courses in statistics, the Ph.D. students had finished all courses in statistics obligatory for students in the psychology Bachelor and Master, and the lecturers could all be expected to have reasonable experience with statistics. The group of undergraduate students consisted of 17 subjects ranging in age from 19 to 24 years old (mean = 21.5, SD = 1.7; 16 women). The group of Ph.D. students consisted of 18 subjects ranging in age from 24 to 37 years old (mean = 27.7, SD = 3.5; 8 women)). The group of lecturers consisted of 16 subjects ranging in age from 30 to 60 years old (mean = 45.8, SD = 9.3; 5 women). The undergraduate students each received € 7 for participating in the experiment.

Stimuli and apparatus. The experiment was conducted on a PC running a program created with MEL 2.0 (Schneider, 1989). On each trial, a short description of a study was presented along with a given *n* and a *p*-value. The *n* was either 10, 50 or 100 and the *p*-value was either .01, .03, .05, .08 or .10. The values of *n* and *p* were combined factorially, resulting in 15 trial types. Each trial type was presented twice, once for the certainty

estimate task and once for the replicability estimate. The resulting 30 trials were presented in random order.

Procedure. On each trial, subjects were asked to give estimates in percentages for either the certainty or replicability probability. It was stressed in the instructions that the exact percentages could not be calculated and that the most reasonable estimate on a scale of 0 to 100 should be given. Only integers could be entered. As soon as the subjects had entered the percentage and confirmed their answer by pressing the 'enter' key on the computer keyboard, the next question appeared. Subjects were prohibited from returning to earlier questions in order to decrease the risk that they would be influenced by previous answers. The task was self-paced and took 20-30 min. After completing the computer task subjects completed a test of basic statistical knowledge which took about 10 minutes.

Two subjects, a university teacher and an undergraduate student, failed to give estimates on all trials. Because an answer was needed in order to continue with the computer program, they filled in a 0 as estimate. Data from these subjects were not analysed. One Ph.D.-student failed to give estimates on all certainty trials, and therefore these certainty data were excluded as well.

5.3 Results

Probability estimates. Figure 5.1, displaying means of estimates for every combination of *n* and *p*-value, shows clear main effects for sample size as well as for *p*-value for both certainty and replicability estimates. This implies that subjects were on average more certain with larger *n*, and also expected on average a higher replicability probability when *n* was larger. On average larger certainty and replicability estimates are given with smaller *p*-values as well. Furthermore, an *n* x *p*-value interaction can be seen for replicability: Whereas for certainty the lines seem to be more or less parallel,

the three lines differ substantially for replicability. This suggests that there seem to be clear differences between the way sample size was interpreted with decreasing *p*-value for replicability.

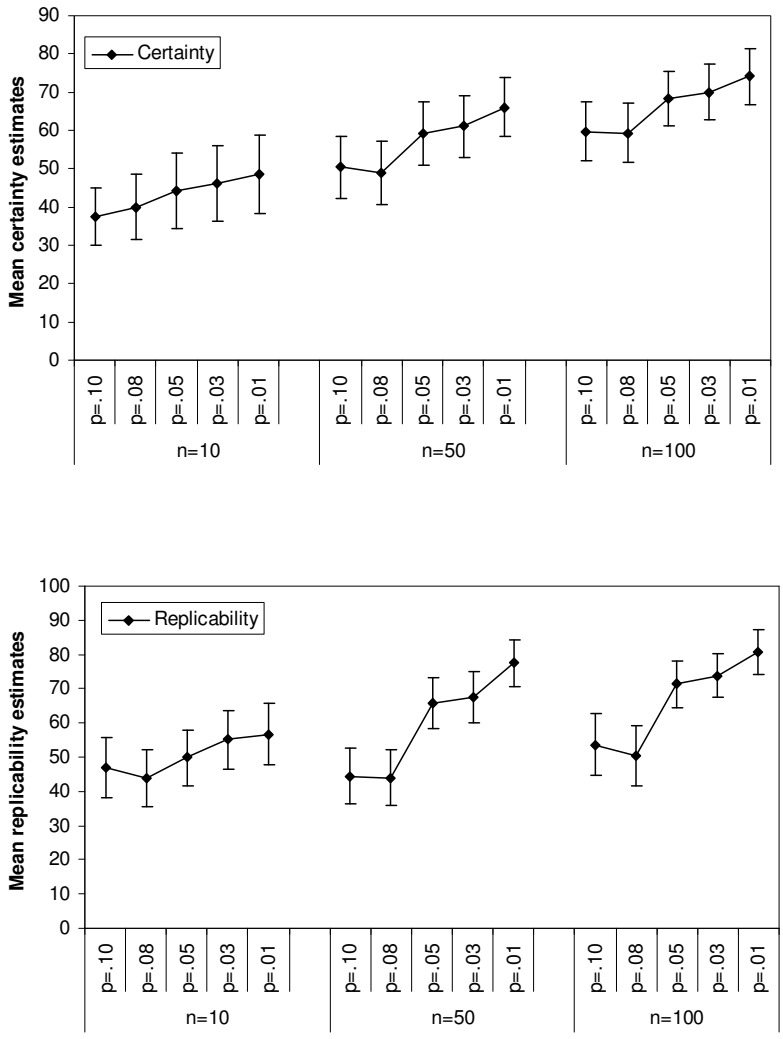

Figure 5.1: Certainty and replicability estimates as a function of n and p-value. Error bars indicate 95% confidence intervals for the mean.

Decreasing *p*-values seem to have a stronger effect for *n*=50 and *n*=100 than for *n*=10, indicating that the subjects seem to be cautious when

interpreting results form small sample sizes. The findings described here are supported by the *F*-tests for a repeated measures ANOVA model, with $p < .0005$ for all effects, except for the *n* x *p*-value interaction for certainty ($p = .36$).

There seem to be discontinuities in the lines of estimates between $p = .05$ and $p = .08$ of Figure 5.1, which seems to reflect a discontinuity in the interpretation of *p*-value due to the conventional use of .05 as an acceptable value. Note, however, that this may be due to the fact that the difference in *p*-values (0.03) is larger than the other differences (all 0.02). On the other hand, it is noticeable that the difference in estimates for these two *p*-values is larger than, for example, the difference between $p = .05$ and $p = .01$ for both probability estimate types. This indicates that subjects may take also the significance (assuming the most common value of .05 as significance level) of the *p*-value into account.

Differences in estimates across subjects. On average, higher certainty and replicability estimates were given for larger sample size, and for smaller *p*-values. This pattern did not, however, hold for every subject. Figure 5.2 shows the interquartile range for probability estimates as a function of *p*-value and *n*. The lengths of the lines between the three points represents the interquartile range (IQR: the range for the middle 50% of estimates) for each of the conditions. The middle point is the median of the condition. The range is an indication of the variation in estimates between subjects for the same condition. The IQRs are in general rather wide (51% on average for certainty, and 49% for replicability), showing that there was large variation as far as estimates are concerned between the subjects.

In order to make inferences from these results, we constructed 95% confidence intervals (CIs) around these interquartile ranges by means of the bootstrap method (Efron & Tibshirani, 1993). The 30 resulting CIs had lower bounds for the IQR ranging from 15% to 44%. Because these are only

the lower bounds of the CIs, it can be safely concluded that the true values for the IQRs are in most cases larger than these lower bounds.

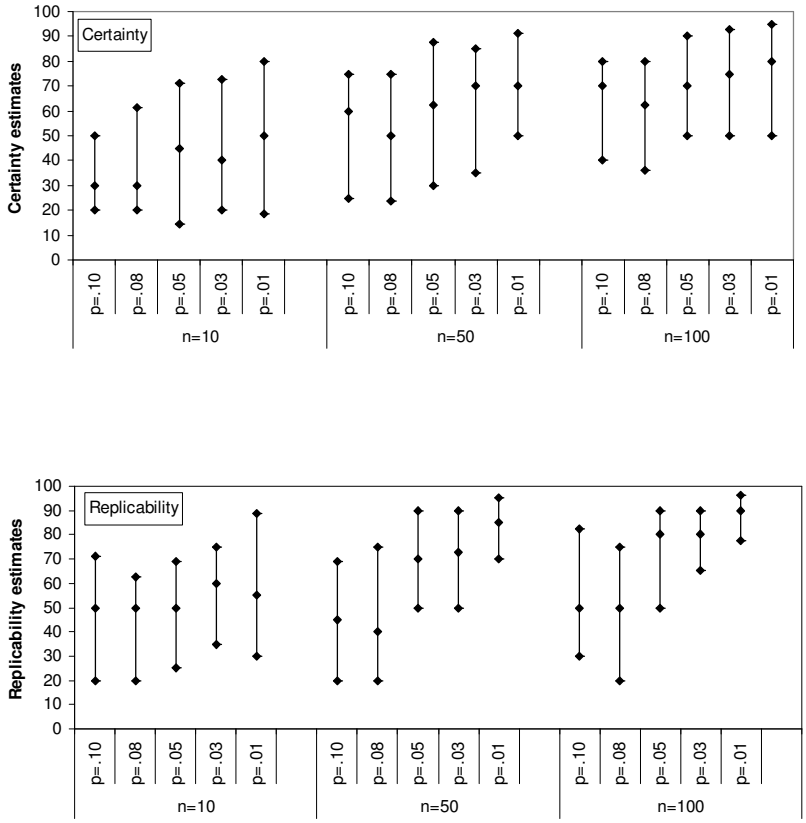

Figure 5.2: Interquartile ranges for certainty and replicability probability estimates as a function of n and p-value.

Patterns of estimates within subjects. To gain insight into within subject variability, we constructed triads of probability estimates with either the same *p*-value and varying *n*, or the same *n* and varying *p*-value (using the *p*-values .10, .05, and .01 to cover the complete range, and triads {.10, .08,

.05} and {.05, .03, .01} to inspect more detailed trends[2]). The triads were then classified as increasing, decreasing, flat, or inconsistent, taking a 5% margin into account (e.g., a triad with the values 30%, 50%, and 50% for n = 10, 50, and 100, respectively, would be classified as increasing). Specifically, we defined increasing triads as triads in which estimates increase more than 5% from the first to the third estimate, with the restriction that the second estimate is not more than 5% lower than the first estimate and not more than 5% higher than the third estimate. Decreasing triads were defined similarly, that is, with the third estimate being more than 5% lower than the first estimate, and the second estimate being not more than 5% higher than the first estimate and not more than 5% lower than the third estimate. Triads of estimates that all differed not more than 5% from each other were considered as flat. Triads were considered inconsistent when they were not increasing, decreasing, or flat. For example, a triad with estimates 83%, 85%, 84% is considered a flat triad, whereas a triad with estimates 83%, 90%, 84% is considered inconsistent. The margin of 5% was chosen in order to prevent small differences between two estimates having too much influence on the categorisation of triads.

Table 5.1 gives the patterns for probability estimates for both probability types with increasing n and fixed p-value. It can be seen that a majority (95% CI for the percentage of increasing trends for certainty: 47% - 60%, for replicability 60% - 72%) of those triads showed an increasing trend, with only a relatively small proportion of decreasing and flat triads. Inconsistent estimates were given in at least 12% (95% CI: 8%-17%) of the triads.

[2] For a given n, we decided to limit attention to triads rather than sets of five p-values. A set of five estimates would have given a large increase of possible patterns, and an increase of arbitrary decisions how to categorize these trends. Furthermore, using triads keeps the trends for p-values given a fixed sample size comparable to the trends for sample size for a fixed p-value.

Table 5.1: Percentages of patterns of probability estimates with increasing n, when p-value is fixed.

	Trends:			
Probability:	Monotonic increasing	Monotonic decreasing	Flat estimates	Inconsistent estimates
Certainty	54%	6%	19%	21%
Replicability	66%	8%	14%	12%

Table 5.2 gives the percentages of triads of each type for both certainty and replicability probability estimates as a function of decreasing *p*-value. A majority of triads showed an increasing trend (95% CI for the percentage of increasing trends for certainty: 46% - 63%, for replicability 58% - 73%).

Table 5.2: Percentages of patterns of probability estimates with decreasing p, when sample size is fixed. P-values of .01, .05 and .10 are compared.

	Trends			
Probability	Monotonic increasing	Monotonic decreasing	Flat estimates	Inconsistent estimates
Certainty	55%	7%	14%	24%
Replicability	66%	6%	12%	16%

The tables for triads of respectively the triads of *p*-values {.01, .03, .05} and {.05, .08, .10} (see Tables 5.3 and 5.4) show that, as might be expected, much more variation in the type of triads can be seen when the spread of the values of the triads is smaller. Although the proportion of monotonically increasing triads is still much larger than the proportion of monotonically decreasing triads, the results indicate that people seem to have difficulties to make clear distinctions between *p*-values that are close to one another. This can also be seen when considering the increase of inconsistent

estimates when compared to the proportion of inconsistent estimates in Table 5.2 (95% CIs for increase of inconsistent certainty estimates for *p*-values {.10, .08, .05} compared to, respectively, *p*-values {.05, .03, .01} and *p*-values {.10, .08, .05} are 15% - 36% and 5% - 25%, and for replicability these are 9% - 28% for both comparisons). Note that this increase is found despite the 5%-margin we used.

Table 5.3: Percentages of patterns of probability estimates with decreasing p, when sample size is fixed. P-values of .01, .03 and .05 are compared.

	Trends			
Probability	Monotonic increasing	Monotonic decreasing	Flat estimates	Inconsistent estimates
Certainty	22%	10%	25%	44%
Replicability	33%	5%	37%	34%

Table 5.4: Percentages of patterns of probability estimates with decreasing p, when sample size is fixed. P-values of .05, .08 and .10 are compared.

	Trends			
Probability	Monotonic increasing	Monotonic decreasing	Flat estimates	Inconsistent estimates
Certainty	35%	11%	22%	33%
Replicability	26%	10%	30%	35%

Most subjects' estimates increased when *p*-values decreased, as can be seen in Table 5.2. Apparently, the way both probabilities depend on *p*-values is intuitively clear to many researchers. Although this might not be a very surprising finding, this result shows that many researchers and students *are* capable of distinguishing estimates based on *p*-values correctly (at least as far as direction is concerned, given that estimates for smaller *p*-values

should result in smaller certainty and replicability estimates), and that the task in our study was not impossible to perform.

The estimates for the effect of sample size were a little less straightforward, but still interpretable. The majority of subjects expected both certainty and replicability to increase with increasing *n*. This is in disagreement with the outcomes of our numerical assessments for both the certainty and the replicability probability. In both cases, we found that the estimates should be almost independent of *n.* (The biggest difference was 3.2%, hence clearly below the 5% margin) In our study, however, only 19% (95% CI: 14% - 24%) of the certainty triads and 14% (95% CI: 10%-19%) were flat, and that within a margin of 5%.

5.4 Discussion

We asked psychology undergraduates, Ph.D. students and lecturers to estimate two probabilities on two questions highly relevant for interpreting research outcomes, given varying values for *n* and the *p*-value. The data were characterized by large differences between subjects, which suggests that interpretations of certainty and replicability will differ. The finding of high intersubject variability could also have been partly due to the difficulty of the task. This supposed difficulty was supported by the high proportion of subjects complaining afterwards about the difficulty of the task.

Because a researcher's interpretation of the outcome of a study should at least partly depend on the two probability estimates discussed here, our findings seem somewhat disturbing for scientific practice. It is possible that, in practice, researchers sometimes consider other factors, such as effect size and confidence intervals, than only the *n* and *p*-value, which were the only statistics given in our experiment. As stated in the introduction, however, we know from previous research that effect sizes (including

means) are given little consideration, and we therefore regard it unlikely that adding effect size would lead to radically different results. We chose not to present confidence intervals in our experiment because they are, unfortunately, seldom given in reporting practice.

We found that, in general, the subjects' estimates for the two probabilities increased with increasing *n* and fixed *p*. A possible explanation for this is that subjects use *n* as a measure of the reliability of the study, and take this into consideration when interpreting the *p*-value. In doing this, however, they ignore the fact that the *p*-value already takes the value of *n* into account.

5.5 Appendix

Numerical assessments were made to determine how certainty and replicability change when *p*-values decrease (while *n* is fixed), or when *n* increases (while *p*-values are fixed). Earlier studies on this topic, using other assumptions, are described at the end of this appendix. When the population standard deviation is known and only the data of one group are of concern, it can easily be proven that both certainty and replicability are independent of *n* (this proof is available upon request). For the more realistic situation that the population standard deviation is unknown and two groups are compared, we were not able to derive such relations mathematically, and therefore we resorted to numerical assessments of such probabilities for various combinations of *n* and *p*-values. In these numerical assessments, we considered the differences between two population means as a stochastic variable with a particular prior distribution. Furthermore, we only considered cases with equal sample sizes (*n*). We also restricted our studies to normally distributed scores in the two populations with the same standard deviation

(σ), and to *t*-tests based on the *t*-statistic $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{1}{n}}\sqrt{s_1^2 + s_2^2}}$, where $\bar{x}_1$ and $\bar{x}_2$ denote the sample means in samples from the two populations, and s_1 and s_2 denote the associated standard deviations.

5.5.1 Certainty

For certainty, we were interested in the following probability: $c(p_g, n) = P(\mu > 0 \mid p = p_g, n)$, with μ being the difference of population means, p_g the *p*-value associated with a pooled sample *t*-test of the observed sample means, and *n* the given sample sizes for both groups. By using Bayes' rule,

$$c(p_g, n) = \frac{P(p = p_g \mid \mu > 0)P(\mu > 0)}{P(p = p_g)}.$$

This, again, can be rewritten as

$$c(p_g, n) = \frac{\left(\int_{m=0}^{\infty} P(p = p_g \mid \mu = m)P(\mu = m \mid \mu > 0)dm \right) P(\mu > 0)}{\int_{m=-\infty}^{\infty} P(p = p_g \mid \mu = m)P(\mu = m)dm}.$$

In order to compute this probability for a given *p*-value, prior information about the distribution of μ is necessary. Because in practice such a prior distribution is rarely available, we decided to use an 'uninformative' prior. That is, in this study, we used symmetric uniform prior distributions for μ, with mean 0 and a range [$-\sigma L$, σL] for a particular large value of *L* (in principle we would like this value to tend to infinity, but in practice we have to resort to real, not too large values). For such symmetric distributions, we have $P(\mu > 0) = \dfrac{1}{2}$. Therefore,

$$P(\mu = m \mid \mu > 0) = \frac{P(\mu = m, \mu > 0)}{P(\mu > 0)} = \begin{cases} \dfrac{P(\mu = m)}{1/2}, & \text{if } m > 0 \\ 0, & \text{if } m \le 0 \end{cases}.$$

Additionally, using that $P(\mu = m)$ is constant on [-σL, σL], we now have

$$c(p_g, n) = \frac{\int_{m=0}^{L\sigma} P(p = p_g \mid \mu = m) P(\mu = m) dm}{\int_{m=-L\sigma}^{L\sigma} P(p = p_g \mid \mu = m) P(\mu = m) dm} = \frac{\int_{m=0}^{L\sigma} P(p = p_g \mid \mu = m) dm}{\int_{m=-L\sigma}^{L\sigma} P(p = p_g \mid \mu = m) dm}$$

.

For every *p*-value and a given *n*, t_g denotes the *t*-value for which it holds that $p_g = P(t > t_g)$. Because every value of p_g is related to only one t_g value, it holds that $P(p = p_g \mid \mu = m) = P(t = t_g \mid \mu = m)$. The distribution for $t \mid \mu = m$ is a noncentral *t*-distribution. In this distribution, $t = \dfrac{d - \mu_0}{s_d}$, with $d = \bar{x}_1 - \bar{x}_2, \mu_0 = 0$ and $s_d = \sqrt{\dfrac{{s_1}^2 + {s_2}^2}{n}}$, and the associated noncentrality parameter is $\delta = \dfrac{\mu - \mu_0}{\sigma_d}$, with $\mu = m$, $\mu_0 = 0$ and $\sigma_d = \sqrt{\dfrac{{\sigma_1}^2 + {\sigma_2}^2}{n}} = \sqrt{\dfrac{2}{n}}\sigma$. The number of degrees of freedom (*df*) is $2n - 2$, just as for the central *t*-distribution. We denote by $p_{nct}(t_g, df, \delta)$ the density for t_g in the noncentral *t*-distribution with *df* degrees of freedom and noncentrality parameter δ. Now we can deduce that

$P(p = p_g \mid \mu = m) = p_{nct}(t_g, 2n-2, \frac{m}{\sqrt{\frac{2}{n}}\sigma})$. To simplify the expressions , we replace $\frac{m}{\sigma}$ by m^* (so, $m = m^*\sigma$), *dm* by σdm^* and the integral limits $-\sigma L$ and σL by $-\sigma$ and σ, respectively for m^*, and thus obtain

$$c(p_g, n) = \frac{\int_{m^*=0}^{L} P(p = p_g \mid \mu = m^*\sigma)dm^*}{\int_{m^*=-L}^{L} P(p = p_g \mid \mu = m^*\sigma)dm^*}.$$

We now searched numerical approximations of $c(p_g, n)$ for various values of n and p_g. Specifically, we approached the integral $\int_{m^*=-L}^{L} P(p = p_g \mid \mu = m^*\sigma)dm^* = \int_{m^*=-L}^{L} P_{nct}(t_g, 2n-2, \frac{m^*}{\sqrt{\frac{2}{n}}})dm^*$ by $\sum_{i=1}^{100000} cp_{nct}(t_g, 2n-2, \frac{m_i^*}{\sqrt{\frac{2}{n}}})$, with *c* being the width of the interval given the number of intervals, and m_i^* the value for m^* for a given *i*, chosen sequentially in the interval [-*L*, *L*] with steps equal to $c = \frac{2L}{100000}$. In our computations we varied *L* as 2, 5 or 10. This procedure was repeated for every combination of *n* and *p*-values, *n* being taking equal to 10, 50, or 100, and *p*-values being taken equal to .10, .08, .05, .03, .01.

The resulting $c(p_g, n)$ values for *L*=10 are displayed in Figure 5.3. For other values of *L*, these figures showed comparable results, provided that *L* was not unreasonably small (defined as, $L < 3$). It can be seen that certainty is essentially independent from *n* (the biggest difference was 0.6%,

for *n*=10 and *n*=100 and *p*-value=.10, for *L*=5), and that certainty increases as *p*-value decreases. Surprisingly, the values of certainty calculated for the same *p*-value and different *n* are lower for larger *n*s, although these differences are negligibly small.

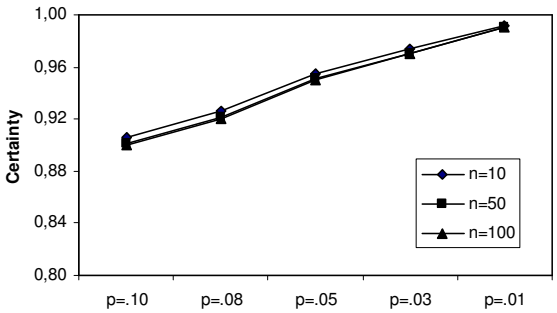

Figure 5.3: Certainty probability as a function of n and p-values, assuming a uniform prior with width 20σ,

5.5.2 Replicability

For a given *n*, we defined replicability as $r(p_g, n) = P(p_{new} < .05 \mid p = p_g, n)$, with p_{new} denoting the *p*-value for a replication study. Using a similar reasoning as for certainty, we get

$$r(p_g, n) = \int_{m=\infty}^{\infty} P(p_{new} < .05 \mid \mu = m) P(\mu = m \mid p = p_g) dm$$

$$= \int_{m=\infty}^{\infty} P(p_{new} < .05 \mid \mu = m) \frac{P(p = p_g \mid \mu = m) P(\mu = m)}{P(p = p_g)} dm$$

$$= \frac{\int_{m=\infty}^{\infty} P(p_{new} < .05 \mid \mu = m) P(p = p_g \mid \mu = m) P(\mu = m) dm}{\int_{m=-\infty}^{\infty} P(p = p_g \mid \mu = m) P(\mu = m) dm}.$$

Again, we assume μ to have a uniform prior distribution on the interval [$-\sigma L$, σL], thus making $P(\mu = m)$ constant, and defining the borders of the interval given by the prior distribution. Thus we get

$$r(p_g, n) = \frac{\int_{m=L\sigma}^{L\sigma} P(p_{new} < .05 \mid \mu = m) P(p = p_g \mid \mu = m) dm}{\int_{m=-L\sigma}^{L\sigma} P(p = p_g \mid \mu = m) dm}.$$

We again have $P(p = p_g \mid \mu = m) = p_{nct}(t_g, 2n-2, \frac{m}{\sqrt{\frac{2}{n}}\sigma})$. This leaves the term $P(p_{new} < .05 \mid \mu = m)$ to be calculated, which can be considered an expression of power, and can be rewritten as

$$P(p_{new} < .05 \mid \mu = m) = \int_{t=t_{.05}}^{\infty} p_{nct}(t_g, 2n-2, \frac{m}{\sqrt{\frac{2}{n}}\sigma}) dt,$$

which is the cumulative density for all *t*-values for which $p < .05$, where $t_{.05}$ is defined by $P(t > t_{.05}) = .05$. Combining these calculations for $P(p = p_g \mid \mu = m)$ and $P(p_{new} < .05 \mid \mu = m)$, replicability can be approximated numerically for every value of *m*. In our simulation study for replicability, we chose again $c = \frac{2L}{100000}$, and *L* = 2, 5 or 10. This procedure was repeated for every combination of *n* and *p*-values, *n* being taking equal to 10, 50, or 100, and *p*-values being taken equal to .10, .08, .05, .03, or .01. The results with *L*=10 are given in Figure 5.4. Again, the results were similar for values of *L* between 3 and 20. It can be seen that replicability, just as certainty, is nearly independent of *n* (the biggest difference was 3.2%, for *n*=10 and *n*=100 and *p*-value=.01, with *L*=10).

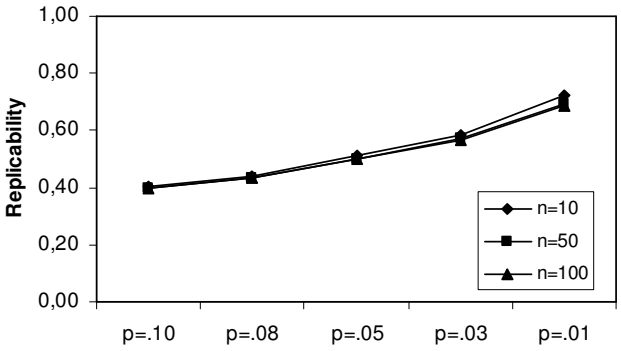

Figure 5.4: Replicability probability as a function of n and p-values, assuming a uniform prior with width 20σ.

5.5.3 Results in the Literature

We found several studies that dealt with the relation between *n* and *p*-value for certainty and replicability, but the studies were conducted in an, in our opinion, incomplete way. In the sequel, some of these studies will be described.

Certainty. As far as certainty is concerned, Bakan (1966) stated that for the same *p*-value one should be more confident with a small *n* than with a large *n*. This finding is in conflict with our findings, and, surprisingly, also contrary to what Rosenthal & Gaito (1963) found, whose data Bakan based his findings on.

Replicability. Greenwald et al. (1996) define replicability as power that a certain found *t*-value would result in a significant finding in an exact replication of this study, making use of Hays' formula (1995) to approximate the power for a non-central *t*-test. They thus showed that replicability is almost independent of *n*. Posavac (2002) tried to improve their approach in a form that is more readily accessible to teachers of statistics, and showed again that replicability is almost independent of *n*. Killeen (2005) introduced the so called p_{rep}, defined as the probability that an effect of the same sign as

that found in the original experiment will be found. This definition is almost identical to our definition of replicability. He showed a direct relation between the classical *p*-value and this p_{rep}. This relation does not include *n*, once again implying that replicability is independent of *n*. Cumming and Maillardet (2006) stated that unless *n* is very small (less than 10), sample size has little effect on replicability.

All the mentioned studies on replicability have in common the finding that sample size has little or no influence on replicability. This contradicts what we found in our calculations because it does seem to depend on *n*, even though only to a very small extent. The difference of our approach with theirs is that they all did not use a prior distribution explicitly, but used the sample outcomes, based on the *p*-value and *n*, as a point estimate for the population parameter. It can be argued that this amounts to using a uniform prior with all its weight focused on one point, and thus in these articles a different, and in our opinion less realistic, prior distribution was used. It can be argued, of course, that it would be even more realistic to base priors on the hypotheses that a researcher has for the study. Such priors need not necessarily follow a uniform distribution. However, given the fact that the simulations were not related to any "real" research question, such priors were impossible to use here. Nevertheless, we think it is fair to state that a uniform prior with a certain width is a more realistic prior distribution than a fixed point estimate.